DOCUMENT RESUME

ED 385 600 TM 024 049

AUTHOR	O'Neill, Kathleen A.; And Others
TITLE	Differential Item Functioning on the Graduate Management Admission Test.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-93-35
PUB DATE	Aug 93
NOTE	90p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	Black Students; Classification; College Entrance Examinations; *Difficulty Level; Higher Education; *Item Bias; Mathematics Tests; *Racial Differences; *Sex Differences; *Test Items; Verbal Tests; White Students
IDENTIFIERS	*Graduate Management Admission Test; *Mantel Haenszel Procedure

ABSTRACT

The purpose of this study was to identify differentially functioning items on operational administrations of the Graduate Management Admission Test (GMAT) through the use of the Mantel-Haenszel statistic. Retrospective analyses of data collected over 3 years are reported for black/white and female/male comparisons for the Verbal and Quantitative Tests. In general, one to six percent of the items were identified as being differentially difficult per comparison with a greater number of items flagged in the female/male analyses than in the black/white analyses. Although the analyses suggested some content characteristics that may be related to differential item functioning, these findings about GMAT items should be considered tentative since only a small number of items was studied, and all investigations were post hoc analyses. Correlations between item difficulty and differential item functioning were generally low, with the exception of quantitative items in the black/white analyses. For these items, a moderately positive relationship existed between item difficulty and the differential item functioning statistic, showing that black examinees performed differentially better than matched whites as item difficulty increased. Eleven tables in the text and four appendixes (one with six tables) provide information on item classification, means and standard deviations, and problem solving and sentence correction items. (Contains 43 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RR-93-35

ED 385 600

RESEARCH REPORT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

H. I. BRAUN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

DIFFERENTIAL ITEM FUNCTIONING ON THE GRADUATE MANAGEMENT ADMISSION TEST

Kathleen A. O'Neill
W. Miles McPeek
Cheryl L. Wild

Educational Testing Service
Princeton, New Jersey
August 1993

BEST COPY AVAILABLE

TM024049

DIFFERENTIAL ITEM FUNCTIONING ON THE GRADUATE MANAGEMENT ADMISSION TEST

Kathleen A. O'Neill, W. Miles McPeek, and Cheryl L. Wild

Copyright © 1993. Educational Testing Service. All rights reserved.

Table of Contents

	Page
List of Tables	2
List of Contents Appendix A	3
List of Contents Appendix B	4
List of Tables Appendix C	5
List of Tables Appendix D	6
Acknowledgments	7
Abstract	8
Introduction	9
Procedures	11
Results	15
Conclusion	36
References	43
Appendix A	A1
Appendix B	B1
Appendix C	C1
Appendix D	D1

LIST OF TABLES

		Page
Table 1	Summary of Sample Sizes by Test Form	13
Table 2	Mean Score Differences between Reference and Focal Groups in Standard Deviation Units	16
Table 3	DIF Means and Standard Deviations for Black and White Groups	18
Table 4	DIF Means and Standard Deviations for Female and Male Groups	19
Table 5	Number and Percent of GMAT Items Flagged in the Black/White Analysis	21
Table 6	Number and Percent of GMAT Items Flagged in the Female/Male Analysis	22
Table 7	Summary of GMAT Variables Studied	24
Table 8	Summary of Variables for Verbal Measure Significant at .05 Level and Beyond	27
Table 9	Summary of Variables for Quantitative Measure Significant at .05 Level and Beyond	30
Table 10	Correlation of Delta and DIF by Test Form for the Black/White Analysis	34
Table 11	Correlation of Delta and DIF by Test Form for the Female/Male Analysis	35

List of Contents APPENDIX A - Item Classifications

	Page
Introduction	A1
Summary of Variables on Which Items Were Classified	A2
Variables Common to All Item Types	A4
Variables Studied in Verbal Measure: Reading Comprehension	A5
Variables Studied in Verbal Measure: Sentence Correction	A9
Variables Studied in Quantitative Items	A12

List of Contents APPENDIX B - Problem-Solving Items

	Page
Problem-Solving Items Flagged in Black/White Analyses	B1
Problem-Solving Items Flagged in Female/Male Analyses	B3

List of Tables APPENDIX C

Table C1 Means and Standard Deviations of Categories for Variables on GMAT Verbal Tests C1

Table C2 Means and Standard Deviations of Categories for Variables on GMAT Verbal Tests: Reading Comprehension C2

Table C3 Means and Standard Deviations of Categories for Variables on GMAT Verbal Tests: Sentence Correction C7

Table C4 Means and Standard Deviations of Categories for Variables on GMAT Quantitative Tests C12

Table C5 Means and Standard Deviations of Categories for Variables on GMAT Quantitative Tests: Problem-Solving C17

Table C6 Means and Standard Deviations of Categories for Variables on GMAT Quantitative Tests: Data Sufficiency C18

List of Contents APPENDIX D - Sentence Correction Items

Page

Sample Sentence Correction Items D1

Acknowledgments

The authors would like to thank the numerous individuals who contributed to the design and implementation of this study. Special thanks should go to Ruth Austria, without whose help we could not have completed this project. Ruth coordinated the coding of the item classifications, keying of the data, organization of the output, preparation of the tables, proofing of the report, and numerous other administrative activities.

Paul Holland also made significant contributions to this report by suggesting the data analyses, doing the basic theoretical work to allow us to use the Mantel-Haenszel statistic, reviewing the data analyses, and reviewing the report. Minhwei Wang and Dorothy Thayer conducted the data analyses. Many test developers were involved in developing hypotheses and coding the data, including Richard Adams, Peter Cooper, Dennis Eglewski, Diane Eisenberg, and Charlotte Solomon. Thanks are also due to the secretaries, Ruthetta Halbower, Evelyn Kalapos, and Sandra Sherman, who spent hours typing the text and the tables of this report. Finally, the Program Research Planning Council deserves our gratitude; without their support, this research would not have been possible. Numerous other individuals were consulted at various steps of this project and, although not listed here, they deserve our thanks.

Abstract

The purpose of this study was to identify differentially functioning items on operational administrations of the Graduate Management Admission Test (GMAT) through the use of the Mantel-Haenszel statistic. Retrospective analyses of data collected over a three-year period are reported for Black/White and female/male comparisons. Two tests are included in the analyses--Verbal and Quantitative. Data concerning the number of items identified as differentially difficult, the characteristics of the test questions associated with differential item difficulty, and the relationship of item difficulty to differential item functioning are discussed. In general, one percent to six percent of the items were identified as being differentially difficult per comparison with a greater number of items flagged in the female/male analyses than in the Black/White analyses. There was little overlap among the items flagged in the two analyses. Although the analyses suggested some content characteristics that may be related to differential item functioning, these findings about the GMAT items should be considered tentative since only a small number of items was studied and all investigations were post hoc analyses. Correlations between item difficulty and differential item functioning were generally low, with the exception of Quantitative items in the Black/White analyses. For these items, a moderately positive relationship existed between item difficulty and the differential item functioning statistic, showing that Black examinees performed differentially better than matched White examinees as item difficulty increased.

Differential Item Functioning on the Graduate Management Admission Test

Differential item functioning (DIF) is said to exist when two groups of people, matched in terms of their relevant knowledge and skill, perform differently on an item. In the summer of 1986, ETS began to introduce a statistical measure of DIF as an additional routine tool in the test development process. The introduction of this tool was the culmination of a series of studies to identify a statistical methodology that could be used in a variety of testing programs to detect test questions exhibiting differential item functioning among various subgroups. This paper presents results of one of the studies--the study investigating the use of the Mantel-Haenszel statistic on retrospective data from Graduate Management Admission Test (GMAT).

A coordinated research program, begun in 1983, evaluated the Mantel-Haenszel statistic as a method to investigate DIF. Papers by Holland (1985), Holland and Thayer (1986), and Phillips and Holland (1986) present the statistical evaluations and development of the procedures. The procedures were then used in retrospective analyses of the NTE Core Battery tests, the Graduate Record Examinations General Test, the Graduate Management Admission Test, the Multistate Insurance Licensing Test, and the Scholastic Aptitude Test. Parts of these retrospective studies have been reported in several places (Carlton & Harris, 1989; McPeek & Wild, 1986, 1987; O'Neill, Wild, & McPeek, 1989; Wild & McPeek, 1986).

Numerous research projects at ETS have investigated differential item functioning over the years. These include applications of delta plot, analyses of variance, standardization, and item response theory techniques

(e.g., Angoff & Ford, 1973; Cardall & Coffman, 1964; Dorans & Kulick, 1986; Lord, 1980). The earlier studies differ from the effort reported here mainly in the statistical procedure used (the Mantel-Haenszel) and the focus on developing procedures that can be used operationally. Operational procedures that are currently in use at ETS are described by Zieky (1988).

Although originally used by biostatisticians (Mantel and Haenszel, 1959), the Mantel-Haenszel (MH) procedure was proposed by Holland as a statistically powerful and efficient index for identifying differentially functioning test items in operational test forms (Holland, 1985). The MH technique involves matching candidates on an appropriate criterion and then calculating an odds ratio based on the performance of the two groups under study. The MH procedure has been compared to procedures intended to identify differential item performance (Holland and Thayer, 1986; Lord, 1980; Marascuilo and Slaughter, 1982; Mellenberg, 1982; Scheuneman, 1979). It has been recommended instead of item response theory (IRT) methods because it is independent of the assumptions underlying IRT models, it is applicable to small samples, and it is inexpensive.

The study in this report investigated differential item functioning in operational administrations of the Graduate Management Admission Test through the use of the Mantel-Haenszel procedure. Differential item functioning is said to exist when two groups of test-takers, matched in terms of their relevant knowledge or skill, perform differently on a test item. A high DIF value may reflect real differences between the groups on what is being tested or it may reflect characteristics of the test item which could be unrelated to the concept being tested. It is the latter type of items which are of concern and which testing programs try not to develop.

The GMAT is designed to help graduate schools of business assess the qualifications of applicants for advanced study in business and management. The GMAT measures general verbal and mathematical skills that are developed over a long period of time and that are associated with success in the first year of study at graduate schools of management. The quantitative sections of the test measure basic mathematical skills and understanding of elementary concepts, and the ability to reason quantitatively, solve quantitative problems, and interpret graphic data. The verbal sections of the test measure the ability to understand and evaluate what is read and to recognize basic conventions of standard written English. The present study is a retrospective look at data collected over a three-year period, 1983-1985, at operational test administrations.

This study used these data to focus on three basic research questions:

1) What percentage of questions is identified as differentially difficult?

2) What are some of the characteristics of the test questions that are associated with differential difficulty?

3) Is there a significant correlation between item difficulty and DIF?

Procedures

Three forms of the GMAT test were analyzed for this study; each form consisted of 85 Verbal items and 65 Quantitative items. The subparts of the Verbal test were reading comprehension (25 items), sentence correction (25 items), and analysis of situations (35 items). Since the analysis of situations items are no longer included in the GMAT, these items were not analyzed in depth; however, they were included in the total verbal score that was used as the criterion to establish the matched groups. The subparts of

the Quantitative test were problem solving (40 items) and data sufficiency (25 items).

In this study, a focal group consisting of Black examinees was compared to a reference group of White examinees matched in ability, and a focal group consisting of women was compared to a reference group of men matched on ability. Table 1 summarizes the sample sizes and administration dates for the three test forms that were analyzed. The subjects in these samples were all United States citizens for whom English was their best language.

As previously noted, in the MH procedure groups of examinees are matched on a relevant criterion, which is often the total score on all available items testing the same skill or knowledge. For the Verbal GMAT items, the total verbal score was the criterion. For the Quantitative items, the total Quantitative score was the criterion. For each measure, examinees at each score level are analyzed separately and the results summarized across score levels (Holland and Thayer, 1986). Relatively unidimensional scores provide better matching of the two groups and are therefore preferable. When the term "matched groups" is used in this report, it is used only in the way described above.

After the MH ratio was calculated, it was then transformed to a scale more familiar to ETS test developers--the difference between the item difficulties of the two matched groups in delta units (deltas have a mean of 13 and a standard deviation of 4, with more difficult items corresponding to larger numbers). We have called this index of differential item difficulty "DIF" in the tables and text. In operational test development practice, a difference in equated deltas greater than one would be very unusual from repeated administrations of the same question. Thus, an absolute value of DIF

Table 1

Summary of GMAT Sample Sizes by Test Form*

Form	Administration Date	White Group	Black Group	Male Group	Female Group
1	10/84	22,774	1,596	20,214	12,297
2	1/85	18,243	1,213	15,585	9,315
3	6/85	17,741	1,360	14,798	9,508

* The total number of examinees in the Black/White analysis is less than the total number in the Female/Male analysis because some examinees belong to other ethnic groups and because other examinees do not grid any designation for race or ethnic group.

equal to or greater than one is used as the flagging criterion. (This corresponds to a difference between groups in percent correct of approximately 10 for questions with percent correct values between 25 and 75.) Since these GMAT tests were analyzed before a reliable procedure for estimating the standard error of the DIF values was established, large sample sizes were used to ensure that even very small differences were statistically significant.

The flagged questions are separated in the summary tables into those questions on which reference group test-takers (White or male examinees) perform better than the matched focal group test-takers (Black or female examinees) with comparable test scores, and those questions for which the reverse held. Summaries are presented by skills or area of knowledge to facilitate interpretation.

For each of the two tests (Verbal and Quantitative), a set of approximately 30 variables was hypothesized that might result in differential performance by group. The items were classified according to each variable, and F-tests were performed to determine if there were statistically significant differences in mean DIF values for the categories within each variable[1]. For example, "minority stimulus" was one of the hypothesized variables. Items were categorized as follows:

(1) Stimulus material refers to Black people
(2) Stimulus material refers to Hispanic people
(3) Stimulus material refers to other minorities
(4) Stimulus material refers to people of no specified ethnic origin
(5) Stimulus material does not refer to people

A listing of all the hypothesized variables and the coding procedures is

---

[1]Standard deviations for the two groups are similar for most variables in this study. Although the F-test is affected by unequal standard deviations, it is fairly robust to violations of heteroscedasticity.

presented in Appendix A. Because items were classified on about thirty variables, many within-variable categories contained few or no items, and there was great overlap among some variables. For example, in the Quantitative measure, the variable "Stimulus Format II: Letters for Mathematical Quantities" refers to the use of *x* or *y* in place of a quantity. This variable overlaps considerably with the algebra category in the primary content variable since algebra often uses letters for mathematical quantities.

In this paper we discuss variables that have significant differences at the .01 level and that have at least five items in the relevant categories.

Results

The results section begins with a brief overview of the data and then presents results specific to the basic questions of interest.

Overview

As shown in Table 1, the Black/White and female/male analyses are based on relatively large sample sizes; other minority groups did not have a sufficient number of examinees to allow DIF analyses to be conducted. Since the DIF values are unstable across samples with small sample sizes and since procedures at the time of data analysis did not permit a precise estimate of the standard error of DIF values, only results involving sufficiently large sample sizes are included. For these analyses, the focal group contained a minimum of 500 examinees.

Table 2 presents the mean Verbal and Quantitative score differences in standard deviation units for the groups being studied. The mean score differences are largest for the Black/White comparisons, with White test-takers having higher mean scores than Black test takers on both Verbal and

Table 2

Mean Score Differences* Between Reference and Focal Groups in Standard Deviation Units**

Black/White Analysis

	Maximum Number of Items***	Form 1	Form 2	Form 3
Verbal	85	-1.1	-1.0	-1.2
Quantitative	65	-1.1	-1.1	-1.2

Female/Male Analysis

	Maximum Number of Items***	Form 1	Form 2	Form 3
Verbal	85	.17	.14	.12
Quantitative	65	-.43	-.42	-.46

---

* Calculated by the formula:

$$\frac{\text{(Mean Score for Focal Group - Mean Score for Reference Group)}}{\text{Total Group Standard Deviation}}$$

** In the Black/White analysis, Blacks are the focal group; in the female/male analysis, females are the focal group.

*** Occasionally an item is dropped from scoring.

Quantitative tests. The female/male standardized mean differences are closer to zero than are those for the Black/White comparisons. A consistent difference in performance is clear: male test-takers have higher mean scores on the Quantitative test; female test-takers have slightly higher mean scores on the Verbal test.

Tables 3 and 4 present the average DIF values, that is, the mean differences between the reference and focal groups when examinees being evaluated for a specific skill or topic are matched on total score measuring the same skill or subject area. For the total Verbal and the total Quantitative score, the expected value is zero and the results may differ from this value only due to the small effects associated with the nonlinearity of the Mantel-Haenszel weighting procedure for combining data across score levels. However, the item types within each of these scores may differ from zero in ways that reflect differences in the groups' performance. As can be seen, these mean values differ slightly from zero only due to small effects associated with the nonlinearity of the Mantel-Haenszel weighing procedure for combining data across score levels. The item types do differ from zero. As Table 3 shows, there are no large differences in performance on the different sections of the test for Black and White examinees although Black examinees score slightly higher than White examinees with comparable test scores on problem-solving questions for two of the three test forms. On the other hand, the female/male comparison reveals larger and more systematic differences. Male examinees perform better than a matched group of female examinees on reading comprehension items for two of the three forms and, to a lesser extent, on problem-solving items; but female examinees perform better than a matched group of male examinees on sentence correction items and on data

Table 3

DIF Means* and Standard Deviations for Black and White Groups

		Form 1	Form 2	Form 3
Verbal				
TOTAL	Mean	.00	-.01	.00
	SD	.33	.38	.47
Reading Comprehension (25 items)	Mean	-.02**	-.08	.03
	SD	.32	.50	.50
Sentence Correction (25 items)	Mean	.06**	.03	.01
	SD	.40	.23	.39
Analysis of Situations (35 items)	Mean	-.02.	.02	-.04
	SD	.30	.35	.51
Quantitative				
TOTAL	Mean	.04	.03	.07
	SD	.49	.46	.56
Problem Solving (40 items)	Mean	.10	.01	.08
	SD	.51	.51	.58
Data Sufficiency (25 items)	Mean	-.05	.06	.04
	SD	.46	.36	.53

* Negative values indicate that White examinees perform better than Black examinees with comparable test scores; positive values indicate the reverse.

** One item not scored.

Table 4

DIF Means* and Standard Deviations for Female and Male Groups

		Form 1	Form 2	Form 3
Verbal				
TOTAL	Mean	.01	.00	.01
	SD	.37	.37	.42
Reading Comprehension (25 items)	Mean	.07**	-.27	-.21
	SD	.41	.33	.41
Sentence Correction (25 items)	Mean	.12**	.12	.29
	SD	.45	.36	.45
Analysis of Situations	Mean	-.12	.11	-.04
	SD	.23	.29	.28
Quantitative				
TOTAL	Mean	.01	.01	.02
	SD	.50	.48	.66
Problem Solving (40 items)	Mean	-.03	-.11	-.10
	SD	.53	.43	.69
Data Sufficiency (25 items)	Mean	.08	.20	.20
	SD	.46	.51	.56

---

* Negative values indicate that male examinees perform better than female examinees with comparable test scores; positive values indicate the reverse.

** One item not scored.

sufficiency items.

Tables 3 and 4 also indicate that the mean DIF values by item type vary from form to form. Since both the test questions and examinees vary from form to form, it is unclear whether differences are due to differences in the sample or to differences in the test content.

Question 1: What percentage of questions are identified as differentially difficult?

Table 5 presents the number and percent of items flagged in the Black/White analyses of the Verbal and Quantitative tests. The table shows that only five percent of the Quantitative items studied had an absolute value of the DIF statistic that exceeded one, and only one percent of the Verbal items studied had an absolute value of the DIF statistic that exceeded one. These totals are not significantly different from what would be identified by chance. As Table 5 also indicates, the problem-solving area exhibits the greatest number of large absolute DIF values: 8 out of the 10 Quantitative items with large DIF values are problem-solving and of these 8 items, Black examinees perform better than a matched group of White examinees on 7 items. The text of the eight problem-solving items is shown in Appendix B.

The results for female and male examinees are shown in Table 6. As in the Black/White analyses, the percentage of items in the female/male analyses whose absolute value of the DIF statistic exceeded one is low: only six percent of the Quantitative items and,only two percent of the Verbal items. Again, these values are only slightly higher than would be expected by chance. Although more items with high DIF values were identified in the female/male analyses than in the Black/White analyses, these numbers are small. Again, 8 of the 12 Quantitative items with large absolute DIF values were problem-

Table 5

Number and Percent of GMAT Items Flagged
in the Black/White Analysis

		Flagged Items with Negative DIF values* DIF < -1			Flagged Items with Positive DIF values DIF > 1			
		Form 1	Form 2	Form 3	Form 1	Form 2	Form 3	Total
Verbal								
Reading Comprehension (74 items)	N	0	0	0	0	1	0	1
	%	0	0	0	0	1	0	1
Sentence Correction (74 items)	N	0	0	0	0	0	0	0
	%	0	0	0	0	0	0	0
Total (148 items)	N	0	0	0	0	1	0	1
	%	0	0	0	0	1	0	1
		Form 1	Form 2	Form 3	Form 1	Form 2	Form 3	Total
Quantitative								
Problem Solving (120 items)	N	1	0	0	2	2	3	8
	%	1	0	0	2	2	3	7
Data Sufficiency (75 items)	N	0	0	0	1	0	1	2
	%	0	0	0	1	0	1	3
Total (195 items)	N	1	0	0	3	2	4	10
	%	1	0	0	2	1	2	5

* Negative items indicate that White examinees perform better than Black examinees with comparable test scores; positive values indicate the reverse.

Table 6

Number and Percent of GMAT Items Flagged
in the Female/Male Analysis

		Flagged Items with Negative DIF values DIF < -1			Flagged Items with Positive DIF values DIF > 1			
Verbal		Form 1	Form 2	Form 3	Form 1	Form 2	Form 3	Total
Reading Comprehension (74 items)	N	0	0	0	0	0	0	0
	%	0	0	0	0	0	0	0
Sentence Correction (74 items)	N	0	0	0	1	0	2	3
	%	0	0	0	1	0	3	4
Total (148 items)	N	0	0	0	1	0	2	3
	%	0	0	0	1	0	1	2
Quantitative		Form 1	Form 2	Form 3	Form 1	Form 2	Form 3	Total
Problem Solving (120 items)	N	1	2	4	0	0	1	8
	%	1	2	3	0	0	1	7
Data Sufficiency (75 items)	N	1	1	0	0	0	2	4
	%	1	1	0	0	0	3	5
Total (195 items)	N	2	3	4	0	0	3	12
	%	2	2	2	0	0	2	6

27

* Negative values indicate that male examinees perform better than female examinees with comparable test scores; positive values indicate the reverse.

-solving items, and of these 8 items, male examinees performed better than a matched group of female examinees on 7 items. The text of these eight problem-solving items is shown in Appendix B; only one problem-solving item flagged in the Black/White analysis was also flagged in the female/male analysis.

Table 6 also shows the number of DIF items with high absolute values for each group by form. Since each GMAT form was administered once, a comparison of DIF values for identical items based on different samples was not possible. This prevents us from analyzing how consistently items with large absolute DIF values are identified in analyzing samples from different test administrations. As Table 6 indicates, the number of flagged items differs slightly across forms; so, too, does the range of DIF values as shown in the frequency distributions provide in Appendix D. For the pooled set of forms, 11 items (3% of the items) received high DIF values in the Black/White comparisons. In the female/male analyses, 15 items (4% of the items) were flagged.

In the analyses that were done, only one item of the 26 flagged items was flagged for both the Black/White and female/male comparisons. In both comparisons this problem-solving item received a negative DIF value indicating that the focal group performed more poorly on the item than did the reference group.

Question 2: What are some of the characteristics of the test questions that are associated with differential difficulty?

Table 7 summarizes the total number of variables studied and the number of variables that contain fewer than five items in all of the categories. For example, in the Quantitative measure, ten variables were hypothesized as possibly related to differential item functioning for all quantitative items.

Table 7

Summary of GMAT Variables Studied

	Number of Variables	Number of Variables with Fewer than 5 Items/Cell
Quantitative Measure		
Quantitative	10	4
Problem Solving	3	3
Data Sufficiency	2	1
Verbal Measure		
Verbal	1	0
Reading Comprehension	12	8
Sentence Correction	10	5

Only six of those variables had at least five items in the categories studied. Variables hypothesized to apply solely to one item type are listed separately.

The full set of significance tables is presented in Appendix C, and significant variables are summarized in Tables 8 and 9 in the main body of this report. (The reader may also want to refer to the detailed descriptions of the variables and classification rules in Appendix A.) A summary of results and preliminary observations about the content characteristics that may be related to DIF values are presented below first for Verbal items and then for Quantitative items. Within each measure, the Black/White results will be discussed before the female/male results. Finally, tentative conclusions will be discussed. These observations and tentative conclusions are intended to suggest further study rather than to be definitive. The results must be interpreted cautiously because of the post hoc nature of the study. In an experimental study, the items could have been chosen to allow the variables to be studied independently and to control for many confounding variables. The tests studied here were not constructed for the purpose of experimental study, and therefore, the variables frequently overlap and the results may be confounded by various other factors.

Only variables with a minimum of five items in each relevant category will be discussed in detail since, in cases with very few items in a cell, one item with an extreme DIF value sometimes will result in significant differences among the categories. Because there was a great deal of overlap among the variables, one item can affect the results for a number of variables.

Verbal Items

Table 8 provides an overview of the results for verbal variables that are significant at the .05 level or beyond. The item type variable shows the results of examinee performance on the two item types: reading comprehension and sentence correction. No significant difference is found between the two item types for Black/White comparisons; however, there is a significant difference in the female/male analyses. Women perform better on sentence correction items than men in the matched reference group; men perform better on reading comprehension items than women in the matched focal group.

All other variables for the Verbal items were studied by separately by item type. For reading comprehension items, the variables with statistical significance for the Black/White comparison were minority stimulus and subject content I. In the minority stimulus category, Black examinees performed better on Black and Hispanic stimulus material than a matched group of White examinees. The converse was true with general material, material that refers to people of no specific ethnic origin.

For the subject content variable, which addresses the overall content of the items, Black examinees performed better on social science content than a matched group of White examinees; on the other hand, White examinees perform better on humanities and science content than a matched group of Black examinees. Given that minority-oriented material can occur more frequently in the social science area than in the other categories, this result is consistent with and confounded with the minority stimulus variable. Further data are necessary to determine whether Black examinees perform better than a matched group of white examinees generally on social studies reading comprehension items or whether this is limited to those social studies

BEST COPY AVAILABLE

Table 8

Summary of Variables for Verbal Measure

Significant at .05 Level and Beyond

		B/W*	F/M*
Item Type		NS	.00
Reading Comprehension:	Minority Reference in Stimulus	.00	NS
Reading Comprehension:	Subject Content I	.00	NS
Reading Comprehension:	Named Gender	NS	.01
Sentence Correction:	Gender Reference	NS	.04
Sentence Correction:	Underline Length-	NS	.04

*B/W = Black/White Analysis, F/M = Female/Male Analysis

items based on Black stimulus material. Although the reading comprehension passages vary in terms of subject matter content of the reading material, the questions are based totally on information presented in the passage. Thus, outside subject matter knowledge is not necessary to answer these questions.

In the female/male analyses of reading comprehension items, one variable proved statistically significant: named gender. On this variable, women performed better than a matched group of men on items in which a mix of genders was named or on items in which no identifiable gender was named than on an item in which a male was named. Since there were no GMAT items in which only a female was named, it was not possible to see if differential group performance occurred on items of this type.

For sentence correction items, no variables achieved statistical significance in the Black/White analyses. In the female/male analyses, both gender reference in the stimulus and underline length were statistically significant. The mean DIF values are positive for all the categories of the gender reference variable, reflecting the overall finding that women perform better on sentence completion items than a matched group of male examinees. The interesting finding about the gender reference variable is that the mean DIF value for items with a male referent in the stimulus is larger than the mean DIF value for other categories and that the mean DIF value for items with both male and female references in the stimulus is smallest.

For underline length, a high positive value occurred in the '8-14 words' and '15 words and over' lengths. In general, the shorter underline units test a single point of grammar or usage and the longer underline units test more complicated structural problems in a sentence (as shown in Appendix D). Results show that the longer underlines have a greater difference in the

performance of the matched groups of men and women. It should be noted that although underline length proved significant, other variables related to the total length of the sentence, the position of the underline, and the element of the sentence being underlined did not prove significant.

Quantitative items

Table 9 summarizes the variables on the Quantitative measure that are significant at the .05 level or beyond. Table 9 includes the variables that were used for all quantitative items as well as those applying only to problem solving or data sufficiency items. In Table 9, eight variables showed statistical significance in the Black/White analyses: gender, real/pure, primary content, quantitative content, the use of letters for mathematical quantities (e.g., x, y), and subject content I.

When Black examinees' performance is compared with a matched group of White examinees on items with a gender reference (items referring to one or more persons), items are differentially more difficult for Black examinees when the reference is female.

Black examinees performed better on pure items than did a matched White examinee group, and the converse was true for real items. The pure items were items based on calculations or formulae, and the real items were word problems which require examinees to set up the problem in order to obtain the correct answer. In general, the real and pure categories of items are of approximately equal difficulty.

For the quantitative content variable, there is a clear distinction on algebraic manipulation where Black examinees score higher than matched White examinees; Black examinees also score slightly higher on ratio/proportion items

Table 9

Summary of Variables for Quantitative Measure

Significant at .05 Level and Beyond

	B/W*	F/M*
Item Type	NS	.00
Gender Reference in Stimulus	.01	NS
Real/Pure	.00	.00
Primary Content Area	.00	NS
Quantitative Content Area	.02	.01
Letters for Mathematical Quantities	.00	NS
Subject Content (Business vs. Other)	.05	NS
Problem Solving: Numbers of Words in Stem	.02	NS
Data Sufficiency: Number of Words in Stimulus	.04	NS

*B/W = Black/White Analysis, F/M = Female/Male Analysis

and factor/multiple items than do matched White examinees.

Items that use letters to represent mathematical quantities such as $x = y + 1$ show higher performance for Black examinees than for the matched group of White examinees. This result was already evident in the Black/White analyses shown in Table 5 where 8 of the 10 flagged items involved this use of letters (and the other 2 items involved powers of integers).

The primary content variable also shows higher Black examinee performance on algebra items than the performance of the matched group of White examinees; however, White examinee performance is higher on geometry and arithmetic items than the performance for the matched group of Black examinees. This result is consistent with the findings on algebraic manipulation and the use of letters for mathematical quantities: Black examinees perform better on algebra-related materials than do White examinees in the comparison group.

The subject content I variable classifies items on whether they contain business or other content. This variable measures whether one group does differentially better on content that is most likely to occur in graduate business programs. For this variable, Black examinees perform less well on business-related materials than a matched group of White examinees.

In the female/male analyses, three quantitative variables proved significant: item type, real/pure, and quantitative content. Item type refers to the two kinds of items in the Quantitative tests: data sufficiency items that have a set of fixed choices, and problem-solving items that have five varying choices. For this variable, women score better on data sufficiency items than a group of men matched on quantitative ability. This result confirms the distribution that was evident in Table 6, which showed fewer negative flagged items for the data sufficiency item type than for the

problem-solving item type.

The real/pure variable, which classifies items as word problems or pure problems that use calculations and formulae, showed that, on average, male examinees scored higher on real items than a matched group of female examinees, and female examinees scored higher on pure items than a matched group of male examinees.

The quantitative content variable shows that female examinees score higher on the categories of algebraic manipulation and problems involving multiples and factors than does a matched group of male examinees. Within the same variable, men perform better on items categorized as ratio/proportion items than do women in a group matched on quantitative ability.

There is only one significant result for the variables unique to problem-solving items: only the number of words in the stem proved significant in the Black/White analyses. For this variable, Black examinees perform better than the comparison group when the stem had less than 50 words; and the converse was true when the stimulus was longer -- 50-99 words. This same variable in data sufficiency items, called number of words in the stem, also proved significant in the Black/White analyses. As in the problem-solving items, Black examinees performed better than did White examinees in a matched group when the stimulus had less than 50 words. For longer stimulus items, White examinees performed better on items where the stimulus had 50-99 words than did a matched group of Black examinees.

The previous discussion of individual variables indicates areas for further investigation. Given that the majority of items are included as part of the analyses of most variables, some of these significant results may be caused by only one or two problem items. Further research is necessary to

isolate which particular variables can be shown to contribute to elevated DIF values.

Question 3: Is there a significant correlation between item difficulty and DIF?

In an effort to determine the relationship between item difficulty (raw delta) and DIF values, correlations were calculated for each test by form. These values are shown in Tables 10 and 11. In the Black/White analyses the correlations for Verbal items are neither strong nor consistent, ranging from -0.28 to 0.17. None of the correlations for Verbal items were statistically significant. For Quantitative items, a moderate positive correlation exists between delta and DIF values for both parts of the Quantitative test. Problem-solving questions have correlations of about 0.4 on all three forms, and data sufficiency questions have a correlation of about 0.6 on two forms (although the third form has a much lower value of 0.11). As Table 10 shows, most of these correlations for the quantitative test were statistically significant at the 0.05 level for a two-tailed test. These correlations show that for Quantitative items in the Black/White analyses, difficult items are associated with high positive DIF values (Black examinees perform better than a matched group of White examinees).

In the female/male analyses, there was generally little relationship found between difficulty and DIF. Although there was a moderate negative correlation between difficulty and DIF values on form 1, this relationship was not evidenced in the other two forms. On the Quantitative tests, neither item type displays any moderate or strong correlation between difficulty and DIF values, and neither item type had a statistically significant correlation on any form.

Table 10

Correlation of Delta and DIF by Test Form for the Black/White Analysis

	No. Items	Test Form 1	Test Form 2	Test Form 3	Pooled Results
Verbal*	85	-0.12	0.08	-0.10	
Reading comprehension	25	-0.07	0.06	-0.28	-0.11
Sentence correction	25	-0.16	0.17	0.10	0.01
Quantitative	65	0.44**	0.28**	0.51**	
Problem solving	40	0.36**	0.35**	0.43**	0.38**
Data sufficiency	25	0.58**	0.11	0.65**	0.48**

*Includes Analysis of Situation items

**Statistically significant at the .05 level (two tailed).

Table 11

Correlation of Delta and DIF by Test Form for the Female/Male Analysis

	No. Items	Test Form 1	Test Form 2	Test Form 3	Pooled Results
Verbal*	85	-0.44**	-0.09	0.01	
Reading Comprehension	25	-0.55**	0.09	-0.06	-0.24
Sentence Correction	25	-0.35	-0.13	0.04	-0.17
Quantitative	65	-0.05	0.14	0.03	
Problem Solving	40	-0.01	0.26	0.08	0.12
Data Sufficiency	25	-0.11	-0.13	-0.12	-0.11

*Includes Analysis of Situation items

**Statistically significant at the .05 level (two tailed).

In addition to correlations between item difficulty and DIF by item type, correlations between these two variables were also obtained by content and format variables. Because many of the variable categories contained only a few items, the correlations for the variables are based on the three test forms pooled together. The overall correlation between difficulty and DIF for reading comprehension was -0.11 for the Black/White analyses and -0.24 for the female/male analyses. For sentence correction items, the overall correlations were 0.01 (Black/White analyses) and -0.17 (female/male analyses). In these analyses, none of the correlations was statistically significant.

For Quantitative items, the correlation between difficulty and DIF on problem-solving items was 0.38 for the Black/White analyses and 0.12 for the female/male analyses. The values for data sufficiency items were 0.48 (Black/White analyses) and -0.11 (female/male analyses). The values for the Quantitative items show that there is a moderately strong relationship between difficulty and DIF values in the Black/White analyses such that the more difficult items are related to positive DIF values; the correlations for both quantitative item types in the Black\White analyses were the only correlations involving variables that were statistically significant (0.05 level, two tailed). The results of the pooled analyses were, thus, consistent with the earlier results obtained separately by form.

Conclusion

Because this study was structured to address three primary issues (the percentage of questions identified as differentially difficult, and the characteristics of these questions and the relationship between item difficulty and DIF), this section will first discuss these areas.

Question 1: What percentage of questions are identified as differentially difficult?

In the Black/White analyses, Table 5 shows that only about three percent of the items were identified as differentially difficult items (one percent of the Verbal items and five percent of the Quantitative items). Of the items identified, the DIF values of all but one item indicated that the focal group (Black examinees) performed better than the matched group (White examinees). In other words, for these GMAT forms, items with the largest absolute DIF values are primarily Quantitative items on which Black examinees score higher than a matched group of White examinees.

For the female/male analyses, the number of items is similar as shown in Table 6: about four percent of the items (two percent of the Verbal items and six percent of the Quantitative items) were identified as differentially difficult. For these items, all of the Verbal items had positive DIF values indicating that the focal group (women) performed better on them, but the large majority of the Quantitative items had negative values indicating that the matched reference group (men) performed better.

These results indicate several things. First, the percentage of items with moderate or large absolute DIF values is low. This finding is consistent with other major testing programs such as the GRE and NTE (McPeek and Wild, in progress). Second, the overall pattern of the flagged items is similar across both focal groups in that there are more Quantitative items flagged for each comparison than there are Verbal items. Third, the flagged items are being drawn from different sections of the DIF distribution. In the Black/White analyses, the focal group performs better than a matched reference group on nearly all of the flagged items whereas in the female/male analyses, the focal group performs less well than a matched reference group on more than half of the flagged items.

When the data are combined across the two analyses, a comparison of the different item types reveals that the average flagging rate for reading comprehension items was 0.7%, for sentence correction items 2.0%, for problem solving items 6.7%, and for data sufficiency items 4.0%. The relatively higher proportion of flagged Quantitative items may suggest that, after being matched on the criterion score, there remain more differences between the groups on the Quantitative items than on the Verbal items. The higher proportion of Quantitative items flagged may be related to content characteristics of the items rather than the surface characteristics which are not directly related to mathematical content (McPeek and Wild, 1987). However, even when examinees are matched on quantitative coursework, differences in performance for matched group of men and women have been found (Doolittle and Cleary, 1987).

Question 2: What are some of the content characteristics of the questions identified as differentially difficult?

Verbal items

Several patterns of significant differences in mean DIF values are evident. Black examinee performance on reading comprehension items is better than that of a matched group of White examinees when the material contains minority references or has social science content. However, these two results are confounded and more research is needed to disentangle these two results. Since minority stimulus material and social science content often appear in the same test items, the results for these two variables are confounded. A similar result has been found for GRE (O'Neill, Wild, and McPeek, in progress) and NTE social science reading items (McPeek and Wild, in progress), although on the GRE items Black examinees performed better than the matched White examinees on both social science and humanities content.

In the female/male analyses of verbal items, men tend to perform better on reading comprehension items than the matched group of women, and women perform better on sentence correction items than did the matched group of men. This differentially superior performance by men on reading comprehension items differs from results on the SAT program (Wendler and Carlton, 1987) which show better performance for women on reading comprehension items than for a matched group of men. However, examinees on both programs are matched on total verbal ability, which, for the SAT includes several item types not contained in the GMAT. Thus, both the samples of examinees and matching criteria vary.

Reading comprehension items including a mix of named genders show better performance by women than a matched group of men. This result may be analogous to the minority stimulus material in the Black/White analyses: when material has a direct interest for the focal group, there may be better performance by the focal group than the matched reference group. This same variable did not prove significant for GRE reading comprehension items, but on NTE reading items a similar result was found: women performed better on items in which a mix of genders was named in comparison to the matched group of men.

In the female/male analyses, the significance of the length of the underline may possibly be attributable to aspects of the content. For example, it is sometimes the case that shorter underlined sections test only a single point or test only a narrow grammatical facet of the sentence. It may be that these aspects of the underline are contributing to the statistical significance. It may also be true that the shorter underlines provide less context for making a determination of the correct response, and this factor is affecting the groups' performance. Further exploration of these issues is indicated.

Quantitative items

In the Black/White analyses, Black examinees perform better than a matched group of White examinees on algebra items. Women also perform better on this content than the matched group of men. This result, which has been found on other testing programs such as the GRE (O'Neill, Wild, and McPeek, in progress) and the ACT Assessment (Doolittle and Cleary, 1987), may reflect the nature and quantity of the mathematics coursework taken by focal group members. The positive mean DIF values on the use of letters for mathematical quantities is consistent with the primary content findings, and has also been found on GRE quantitative items (O'Neill, Wild, and McPeek, in progress). Since algebra items often involve letters used for mathematical quantities for items, the results for these variables are confounded for Blacks and, to some extent, for women matched with the appropriate reference group.

In the Black/White analyses, the length of the stimulus appeared significant for both types of quantitative items. On GRE quantitative items, Black examinees perform better than the matched reference group when the length of the stem was shortest (1-7 words), but not on other lengths (O'Neill, Wild, and McPeek, in progress). One possible explanation for why stimulus length is significant for Black examinees on GMAT Quantitative items, but not on Verbal items, is that the matching criterion for Verbal items is the total Verbal score (which results in matched groups that are roughly equivalent in reading ability) whereas the matching criterion for Quantitative items is the total Quantitative score (which results in groups matched in quantitative ability, not verbal ability).

Black examinees also perform better than the comparison group on items using calculations, equations, or formulae, but worse than the comparison group

on word problems. The reason for this discrepancy is not clear but one area of exploration has been suggested by SAT data gathered on the relationship of the item to the curriculum (Carlton and Harris, 1992). This information is targeted at whether an item is more or less like the problems students have encountered in textbooks and homework assignments. Preliminary results indicate that the focal group outperforms the matched reference group when the item is similar to problems contained in textbooks or homework, but not in other situations. It should also be noted that word problems are longer than pure problems, and there is some confounding of these results with the length of the stimulus. Thus, another explanation of the results might be that since the reading ability of the two examinee groups differs, test questions with a heavier reading load may reflect this difference.

In the female/male analyses of quantitative items, mean DIF values are positive and significant for data sufficiency items, and negative for problem-solving items. This suggests that women perform better than a matched group of men on data sufficiency while the converse is true for problem solving. Other studies (e.g., McPeek and Wild, 1987; and O'Neill, Wild, and McPeek, 1989) have also noted that women perform differentially less well on word problems, even when the women and men have been matched on coursework (Doolittle and Cleary, 1987).

Question 3: Is there a significant correlation between item difficulty and DIF?

In general, the only significant relationship between item difficulty and DIF values appears on Quantitative items, particularly problem-solving items, on which there is a positive correlation between difficulty and DIF values. Black examinees perform relatively better on harder items than on easier items in

comparison with a matched group of White examinees. No relationship was found for the GMAT verbal items, although a correlation has been reported on GRE verbal items (Freedle and Kostin, 1988).

Research Implications

Although analyses suggested some content characteristics that may be related to differential item functioning, the findings should be considered exploratory. Additional data are required to test many of the hypotheses generated for this study. Because there are often only a few items within a category on a single test form, a number of additional test forms will need to be analyzed in order to have enough items to assess fully the relationship between differential item difficulty and single variables or combinations of variables. Once hypotheses that may account for differential item difficulty have been identified, research studies to test these hypotheses will need to be conducted. The experimental manipulation of test questions especially designed to test particular hypotheses will be required. The implications of these findings (both post hoc analyses and experimentally designed analyses) would need to be carefully assessed by both internal and external advisory committees in order to make decisions about the implications that exist for test content.

References

Angoff, W.H. and Ford, S.F. (1973). Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 10, 95-106.

Bleistein, C.A. and Wright, D. (1986, April). Assessment of unexpected differential item difficulty for Asian-American candidates on the Scholastic Aptitude Test. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Breslow, N.E. and Liang, K.Y. (1982). The variance of the Mantel-Haenszel estimator. Biometrics, 38, 943-952.

Cardall, D. and Coffman, W.E. (1964). A method for comparing the performance of different groups on the items in a test (Research Bulletin RB-64-61). Princeton, NJ: Educational Testing Service.

Carlton, S.T. and Harris, A.M. (1989a). Female-male performance difference on SAT: Causes and correlates. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Carlton, S.T. and Harris, A.M. (1989b). Characteristics associated with differential item performance on the SAT: Selected ethnic group comparisons. Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco.

Carlton, S.T. and Harris, A.M. (1992). Characteristics associated with differential item functioning on the Scholastic Aptitude Test: Gender and majority/minority group comparisons (GRE Research Report 92-64). Princeton, NJ: Educational Testing Service

DeMauro, G., & Huffman, R. (1989). Ethnic group performance on a listening skills test. Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco.

Doolittle, A.E. (1983). The reliability of measuring differential item performance. Paper presented at the annual meeting of the American Educational Research Association, Quebec.

Doolittle, A.E. and Cleary, T.A. (1987). Gender-based differential item performance in mathematics achievement items. Journal of Educational Measurement, 24, 157-166.

Dorans, N.J. (1986, April). Two new approaches to assessing unexpected differential item performance: standardization and the Mantel-Haenszel method. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Dorans, N.J. and Kulick, E.M. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. Journal of Educational Measurement, 4, 355-368.

Dwyer, C. A. (1986). Using indices of differential item functioning in test development decisions. Paper presented at the annual meeting of the American Psychological Association, Washington, DC.

Freedle, R. and Kostin, I. (1988). Relationship between item characteristics and an index of differential item functioning (DIF) for four GRE verbal item types (GRE Research Report 85-3). Princeton, NJ: Educational Testing Service

Holland, P.W. (1985). On the study of Differential Item Performance without IRT. Proceedings of the Military Testing Association, October 1985.

Holland, P.W. and Thayer, D.T. (1986). Differential item performance and the Mantel-Haenszel procedure. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Linn, R.L. and Harnish, D.L. (1981). Interactions between item content and group membership on achievement test items. Journal of Educational Measurement, 18, 109-118.

Linn, R.L., Levine, M.V., Hastings, C.N., and Wardrop, J.L. (1981). Item bias in a test of reading comprehension. Applied Psychological Measurement, 5, 159-173.

Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillside, NJ: Erlbaum.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.

Marascuilo, L.A. and Slaughter, R.E. (1981). Statistical procedures for identifying possible sources of item bias based on chi-square statistics. Journal of Educational Measurement, 18, 229-248.

McPeek, W.M. and Wild, C.L. (1986). Performance of the Mantel-Haenszel statistic in a variety of situations. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

McPeek, W.M. and Wild, C.L. (1987) Characteristics of quantitative items that function differently for men and women. Paper presented at the annual meeting of the American Psychological Association, New York.

McPeek, W.M. and Wild, C.L. (1992) Identifying differentially functioning items in the NTE core battery (Research Report 92-62). Princeton, NJ: Educational Testing Service.

Mellenberg, G.J. (1982). Contingency table models for assessing item bias. Journal of Educational Statistics, 7, 105-118.

O'Neill, K.A., Wild, C.L., & McPeek, W.M. (1989). Gender-related differential items performance on graduate admissions tests. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

O'Neill, K.A., Wild, C.L., & McPeek, W.M. (1989). Characteristics of GRE verbal test items that show differential item functioning for Black and White examinees. Paper presented at the National Council of Measurement in Education, San Francisco, CA.

O'Neill, K.A, Wild, C.L., and McPeek, W.M. (In progress) Differential Item Functioning on the Graduate Record Examinations General Test.

Pearlman, M.A. (1987). Trends in women's total score and item performance on verbal measures. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Phillips, A. and Holland, P.W. (1986). A new estimator of the variance of the Mantel-Haenszel log-odds-ratio estimator (Research report no. 86-21). Princeton, NJ: Educational Testing Service.

Rogers, J. and Kulick, E. (1986). An investigation of unexpected differences in item performance between Blacks and Whites taking the SAT. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Scheuneman, J.D. (1979). A method of assessing bias in test items. Journal of Educational Measurement, 16, 143-152.

Schmitt, A.P. (1986, April). Unexpected differential item performance of Hispanic examinees. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Schmitt, A.P. and Dorans, N.J., eds. (1987). Differential Item Functioning on the Scholastic Aptitude Test. Princeton, NJ: Educational Testing Service.

Schmitt, A.P., Bleistein, C.A., and Scheuneman, J.D. (1987). Determinants of differential item functioning for Black examinees on Scholastic Aptitude Test analogy items. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Shepard, L.A., Camilli, G., and Williams, D.M. (1985). Validity of approximation techniques for detecting item bias. Journal of Educational Measurement, 22, 77-105.

Welch, C.J., Ackerman, T.A., and Doolittle, A.E. (1987). An examination of statistical procedures for detecting cross-cultural differential item performance. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Welch, C.J. and Doolittle, A.E. (1988). Gender-based differential item performance in English usage items. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Wendler, C.L.W. and Carlton, S.T. (1987). An examination of SAT verbal items for differential performance by women and men. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Wild, C.L., Durso, R., and Rubin, D.B. (1982). Effect of increased test-taking time on test scores by ethnic group, years out of school, and sex. Journal of Educational Measurement, 19, 19-28.

Wild, C.L. and McPeek, W.M. (1986). Performance of the Mantel-Haenszel statistic in identifying differentially functioning items. Paper presented at the annual meeting of the American Psychological Association, Washington, DC.

Wright, D. (1986). An empirical comparison of the Mantel-Haenszel and standardization methods of detecting differential item performance. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Zieky, M. (1988). Differential item difficulty in test development at Educational Testing Service. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

APPENDIX A

Item Classifications

This appendix presents a detailed description of the classification scheme used to investigate GMAT tests in this research study. These classifications are used to investigate whether specific item characteristics are related to differential item functioning values.

The chart on the following page summarizes the variables that were used by test developers to categorize items in the Verbal and Quantitative measures.

It must be noted that test developers did, in some instances, use the same generic titles as classification variables for different tests, but with different divisions and meanings. Such classification variables have the word varies in parentheses next to them. For example, the classification variable Item Type (varies) has different divisions and different meanings for each of the tests studied. In contrast, four variables—minority stimulus, gender reference in the stimulus, negative item, and Roman numeral format—are the same for all tests to which they apply. Following the chart are the descriptions of how items were classified within each variable for each test. These descriptions will clarify the differences in classification schemes among tests.

Graduate Management Admission Test

Summary of Variables on which Items were Classified

Variables	Verbal		Quantitative
	Reading Comprehension	Sentence Correction	Data Sufficiency and Problem Solving
Minority Stimulus	X	X	X
Gender Reference in Stimulus	X	X	X
Negative Item	X		X (PS only)
Roman Numeral Format	X		X (PS only)
Number of Words in Stimulus (varies)	X	X	X (DS only)
Item Type (varies)	X	X	X
Real/Pure			X
Primary Content Area			X
Multiple Categories			X
Stimulus Format I: Pictures			X
Stimulus Format II: Variables			X
Subject Content I (varies)	X	X	X
Subject Content II	X	X	
Subject Content III	X		
Q Content Area			X
Length of Stimulus			X (DS only)
Data Sufficiency Key			X (DS only)
Question Format	X		
Length of Stem	X		X (PS only)

Graduate Management Admission Test

Variables on which Items were Classified

Variables	Verbal		Quantitative
	Reading Comprehension	Sentence Correction	Data Sufficiency and Problem Solving
Line References	X		
Length of Options	X		
Item Specifications	X		
Named I - Minority	X		
Named II - Gender	X		
Underline Position		X	
Underline Element		X	
Underline Length		X	
Sentence Correction Key		X	
Major Testing Point		X	
Ancillary Testing Points		X	

GMAT VARIABLES COMMON TO ALL ITEM TYPES

MINORITY STIMULUS

1 - Black	Stimulus refers to Black people.
2- Hispanic	Stimulus refers to Hispanic people.
3 - Other Minority	Stimulus refers to other minorities.
4 - General	Stimulus refers to people of no specified ethnic origin.
5 - Nothing	Stimulus does not refer to people.

GENDER REFERENCE IN STIMULUS

1 - Female	Stimulus refers to females.
2 - Male	Stimulus refers to males.
3 - Mixed or Unidentifiable	Stimulus refers to both males and females.
4 - Neutral	Stimulus does not refer to people.

NEGATIVE ITEM (No/Except) (Does not apply to Verbal: Sentence Correction)

1 - Negative Stem	Use "NOT," "CANNOT," "EXCEPT," "LEAST," "INCORRECT," "FALSE," etc., in stem
2 - Positive Stem	Do not use "NOT," "CANNOT," "EXCEPT," "LEAST," "INCORRECT," "FALSE," etc., in stem

ROMAN NUMERAL FORMAT (Does not apply to Verbal: Sentence Correction)

1 - Roman	Involves Roman numeral format
2 - Non-Roman	Does not involve Roman numeral format

GMAT VARIABLES STUDIES IN VERBAL: READING COMPREHENSION

(See also minority stimulus, gender reference in stimulus, negative item, and Roman numeral format in the section "Variables Common to All Item Types.")

NUMBER OF WORDS* IN STIMULUS (Passage) — Refers to the length of the reading passage

1 = Less than 450 words

2 = 450 - 470 words

3 = More than 470 words

ITEM TYPE — Refers to the type of Verbal item classified

1 = Reading Comprehension

2 = Writing - Sentence Correction

QUESTION FORMAT

1 = Open — An open question is one in which the actual question being asked (the stem) is not a complete sentence and is completed by the answer choices (options).

2 = Closed — A closed question is one in which the actual question being asked (the stem) is a complete sentence and the answer choices (options) are independent units (words, phrases, or sentences).

LENGTH OF STEM — Refers to the length of each question (stem) following a reading passage. Each question (item) was coded individually for the number of words*.

1 = Less than 15 words

2 = 15 - 24 words

3 = 25 - 34 words

4 = more than 34 words

*Hyphenated words and numerals were counted as one word for this study.

LINE REFERENCES

1 - Stem directs candidate to a specific line in the passages and lines are given.

2 - Stem directs candidate to a specific paragraph in the passage.

3 - Stem does not direct candidate to any specific part of the passage.

LENGTH OF OPTIONS (A through E)

1 - Less than 20 words*

2 - 20 - 39 words

3 - 40 - 59 words

4 - 60 - 79 words

5 - 80 - 99 words

6 - 100 words or over

SUBJECT CONTENT I	Refers to the overall subject content of the stimulus (reading passage).
1 - Humanities	The passage refers to literature, art, music, philosophy, architecture, dance, and theatre. It is further subdivided in Subject Content II.
2 - Social Studies	The passage refers to history, sociology, psychology, political science, and anthropology.
3 - Sciences	Science content. This is further subdivided in Subject Content III.
4 - Business Related	The passage refers to economics, group dynamics, industrial relations, international trade, manufacturing, marketing, sales, systems theory, management, etc.

*Hyphenated words and numerals were counted as one word for this study.

SUBJECT CONTENT II — Refers to a subdivision of the Humanities

1 - Passage refers to Literature only

2 - Passage refers to all other Humanities

SUBJECT CONTENT III — Refers to a subdivision of the Sciences

1 - Biological Sciences — The passage refers to botany, zoology, genetics, microbiology, etc.

2 - Physical Sciences — The passage refers to chemistry, physics, earth sciences (astronomy, geology, etc.)

ITEM SPECIFICATIONS

1 - Main Idea — Question asks for the explicit or implicit main idea or purpose of the passage.

2 - Supporting Detail — Question asks for identification of details or supporting ideas explicitly stated in the passage.

3 - Inference — Question asks about inferences supported by the passage.

4 - Application — Question asks candidate to apply information found in the passage to a situation not described in the passage.

5 - Logical Structure — Question asks about the logical structure and organization of the passage.

6 - Style — Question asks about the style or tone of the passage or parts of the passage.

NAMED I - MINORITY

Blank - No one named

1 - Black named

2 - Hispanic named

3 - Native American named

4 - Asian American named

NAMED I - MINORITY (cont'd)

5 - Other Minority named

6 - Nonminority named

NAMED II - GENDER

Blank - No one named

1 - Female named

2 - Male named

3 - Mixed or Unidentifiable named

GMAT VARIABLES STUDIED IN
VERBAL: SENTENCE CORRECTION

(See also minority stimulus and gender reference in stimulus in the section "Variables Common to all Item Types.")

NUMBER OF WORDS* IN STIMULUS

1 - Less than 20 words

2 - 20 - 29 words

3 - 30 - 39 words

4 - 40 words and above

ITEM TYPE — Refers to the type of Verbal item classified

1 - Reading Comprehension

2 - Writing - Sentence Correction

UNDERLINE POSITION

1 - Beginning	Includes the first, but not the last word in the stimulus sentence.
2 - Middle	Includes neither the first nor the last word in the stimulus sentence.
3 - End	Includes the last, but not the first word in the stimulus sentence.
4 - Whole	The entire stimulus sentence is underlined.

UNDERLINE ELEMENT

1 - Main Clause Element	Independent clause elements only
2 - Subordinate Element	Subordinate elements only
3 - Main Clause and Subordinate Elements	

*Hyphenated words and numerals were counted as one word for this study.

UNDERLINE LENGTH

1 - 1 - 7 words*

2 - 8 - 14 words

3 - 15 words or more

SENTENCE CORRECTION KEY

1 - Answer choice A, identical to stem

2 - Answer choices B through E

SUBJECT CONTENT I

1 - Business Related	Includes economics, group dynamics, industrial relations, international trade, manufacturing, marketing, sales, systems theory, management, etc.
2 - Domestic and Foreign Affairs	Includes politics exclusive of finance and business
3 - Humanities	Includes literature, art, music, architecture, philosophy, etc. It is further subdivided in Subject Content II.
4 - Science	Includes biological and physical sciences, as well as medicine, physical anthropology and technology.
5 - Social Sciences	Includes history, sociology, psychology, etc.
6 - Other	

SUBJECT CONTENT II — Refers to a subdivision of the Humanities

1 - Refers to Literature only

2 - Refers to all other humanities

*Hyphenated words and numerals were counted as one word for this study.

MAJOR TESTING POINT

Each item was coded for one of the major testing points below:

1 - Agreement	Consistency in person and number, pronoun reference agreement, noun-verb agreement
2 - Construction - Grammatical	Sentence fragments, run-ons, improperly formed constructions
3 - Construction - Rhetorical	Constructions that are awkward, wordy, redundant, imprecise, or unclear, although free of grammatical errors
4 - Diction	Word choice: includes parts of speech, among/between, as/like, mass/count words
5 - Idiom	Prepositions, constructions, correlatives, pronoun usage, forms of comparison
6 - Logical Predication	Modification problems, illogical reference, illogical comparison, distortion of meaning
7 - Negation	Supersedes other major testing point categories
8 - Parallelism	
9 - Verb Form	Other than problems of agreement or parallelism - includes tense, incomplete verb forms

ANCILLARY TESTING POINTS — Refers to testing points subordinate to the major testing point. Zero to 4 ancillary testing points were coded in addition to the major testing point. The same codes used in Major Testing Point (see 1 through 9 above) were used.

GMAT VARIABLES STUDIED IN QUANTITATIVE: DATA SUFFICIENCY AND PROBLEM SOLVING

(See also minority stimulus, gender reference in stimulus, negative item, and Roman numeral format in the section "Variables Common to All Item Types.")

NUMBER OF WORDS IN STIMULUS

1 - Below 50 words

2 - 50 - 99 words

3 - 100 - 149 words

4 - 150 - 200 words

5 - Over 200 words

Hyphenated words, numbers, and mathematical symbols were counted as one word for this study. Figures, graphs, charts and tables were also included in the word counts. For example, Thirty-seven percent of 482 is how much more than 27 percent of 482? would equal 13 words and would be classified as 1.

ITEM TYPE

1 - Data Sufficiency: Fixed Format

Refers to the type of Quantitative item classified

Example: A ladder standing on level ground is leaning against a vertical wall. What is the vertical distance from the top of the ladder to the ground?

(1) The ladder is 50 feet long

(2) The distance from the wall to the foot of the ladder is 25 feet.

Options:

A if statement (1) ALONE is sufficient, but statement (2) alone is not sufficient to answer the question asked;

B if statement (2) ALONE is sufficient, but statement (1) alone is not sufficient to answer the question asked;

C if BOTH statements (1) and (2) TOGETHER are sufficient to answer the question asked, but NEITHER statement ALONE is sufficient;

D if EACH statement ALONE is sufficient to answer the question asked;

ITEM TYPE (cont'd)

E if statements (1) and (2) TOGETHER are NOT sufficient to answer the question asked, and additional data specific to the problem are needed.

2 - Problem Solving: Five-Choice Variable Options

Example: Thirty-seven percent of 482 is how much more than 27 percent of 482?

(A) 0.482 (B) 4.82 (C) 48.2
(D) 173.52 (E) 308.48

REAL/PURE

1 - Real

Includes story or word problems. The candidate must figure out how to set up the problem in order to obtain the correct answer.

2 - Pure

The candidate must use theory, formulae, calculations, etc., to obtain the answer

PRIMARY CONTENT AREA

1 - Arithmetic

2 - Algebra

3 - Geometry

MULTIPLE CATEGORIES

1 - Problems that can be solved using only arithmetic, algebra or geometry.

2 - Problems that can be solved using arithmetic and/or algebra.

3 - Problems that can be solved using arithmetic and/or geometry.

4 - Problems that can be solved using algebra and/or geometry.

STIMULUS FORMAT I: PICTURES

1 - Figure	Picture does not have a coordinate system (has a triangle, square, etc.).
2 - Graph	Picture has a coordinate system, or is a line, bar, or circle graph.
3 - Table	Picture has data presented in rows and columns.
4 - None	Pictures are not present in the stem.

STIMULUS FORMAT II: VARIABLES

1 - Variables in options

2 - Variables in stem

3 - Variables in options and stem

4 - None - Variables are not present in options or stem.

SUBJECT CONTENT I

1 - Business Related	Includes economics, group dynamics, industrial relations, international trade, manufacturing, marketing, sales, systems theory, management, etc.
2 - Other	

Q CONTENT AREA

1 - Algebraic manipulations	Problems which are solved by simplifying algebraic expressions, squaring algebraic expressions, combining the terms, using the distributive property, etc.
2 - Ratio/proportion	Problems which involve comparing two quantities by means of division, determining the equalities of two ratios, etc.
3 - Factor/multiple	Problems which involve multiplying quantities. Look for words like divisor or divisible in the problem.
4 - None of the above	

DATA SUFFICIENCY KEY

See example given under ITEM TYPE: Data Sufficiency for a more complete description of the options A through E.

1 - A option

2 - B option

3 - C option

4 - D option

5 - E option

Appendix B

Problem-Solving Items Flagged in DIF Analyses

Items Flagged in Black/White Analyses

If $3x - 4y = 17$ and $2x - 5y = 6$, then $x + y =$

(A) -11 (B) $\frac{37}{14}$ (C) $\frac{119}{14}$ (D) 11 (E) 23

DIF: +1.19

17. $\sqrt{\left(\frac{\sqrt{2}}{2} + \frac{\sqrt{2}}{2}\right)^2 + \left(-\frac{\sqrt{2}}{2} - \frac{\sqrt{2}}{2}\right)^2} =$

(A) 0 (B) $\frac{1}{2}\sqrt{2}$ (C) $\sqrt{2}$ (D) 2 (E) $2\sqrt{2}$

DIF: +1.43

20. If $\frac{x}{y} = \frac{y+1}{x+1}$ and x is not equal to y, what is the value of y in terms of x?

(A) $-x - 1$ (B) $-x$ (C) $-x + 1$
(D) $x - 1$ (E) $x + 1$

DIF: +1.63

T2. In the XY-plane, if $\left(k, -\frac{k}{2}\right)$ is a point on the graph of $3x - 2y = 8$, then $k =$

(A) -4
(B) -2
(C) 2
(D) 4
(E) 8

DIF: +1.00

Items Flagged in Black/White Analyses

6. A trillion is written as the digit 1 followed by 12 zeros. Which of the following is equal to a trillion?

(A) $10^{13} \times 10$ (B) 10^{13} (C) $10^{10} \times 10^{2}$
(D) 10^{11} (E) $10^{9} \times 10$

DIF: +1.15

18. If ${}_kS_n$ is defined to be the product $(n + k)(n - k + 1)$ for all positive integers k and n, which of the following expressions represents ${}_{k-1}S_{n+1}$?

(A) $(n + k)(n - k + 2)$
(B) $(n + k)(n - k + 3)$
(C) $(n + k + 1)(n - k + 2)$
(D) $(n + k + 2)(n - k + 1)$
(E) $(n + k + 2)(n - k + 3)$

DIF: +1.57

14. If x pencils cost y dollars, then at this rate the number of dollars that y pencils cost is

(A) x (B) $\frac{x^2}{y^2}$ (C) $\frac{y^2}{x}$ (D) $\frac{x^2}{y}$ (E) y

DIF: +1.10

5. Car X averages 18 miles per gallon in city driving and 27 miles per gallon in highway driving. If the fuel tank capacity is 25 gallons, how many more miles per tankful of fuel can be expected in highway driving than in city driving?

(A) 562.5 (B) 225 (C) 34 (D) 22.5 (E) 9

DIF: -1.28
DIF: -1.85(Female/Male)

BEST COPY AVAILABLE

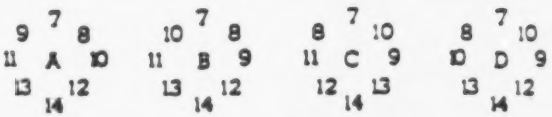
Items Flagged in Female/Male Analyses

4. An airplane leaves the deck of a carrier and heads due west at an average rate of 300 miles per hour. The carrier travels in the same direction at an average rate of 30 miles per hour. *If the airplane* radio has a range of 450 miles, for how many hours can the plane maintain radio communication with the carrier?

(A) $\frac{3}{5}$ hr. (B) $1\frac{4}{11}$ hr. (C) $1\frac{2}{3}$ hr.

(D) 2 hr. (E) $2\frac{1}{3}$ hr.

DIF: -1.53



6. Which of the circles of numbers above does NOT have the property that each number in it has a difference of at most 3 from each of the numbers next to it in the circle?

(A) A (B) B (C) C (D) D

(E) All the circles have the property.

DIF: +1.00

7. If a column of water one foot high exerts a downward pressure of 0.433 pound per square inch, approximately what is the height, in feet, of a column that exerts a downward pressure of 87 pounds per square inch?

(A) 380 (B) 200 (C) 38 (D) 20 (E) 2

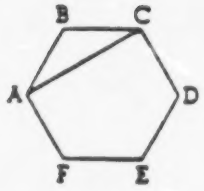
DIF: -1.45

14. A certain driver's reaction time before ving the brakes is 1 second and the distance : .uto-mobile travels before coming to a stop after the brakes are applied is 30 feet at 20 miles per hour (1 mile = 5,280 feet). The total stopping distance of this driver at 20 miles per hour is approximatelv

(A) 30 ft.
(B) 60 ft.
(C) 90 ft.
(D) 1,760 ft.
(E) 1,790 ft.

DIF: -2.11

Items Flagged in Female/Male Analyses

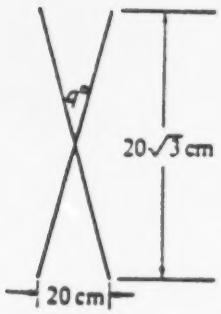


16. The figure above is a regular hexagon. If each side has length 10, what is the length of AC ?

(A) 10 (B) $10\sqrt{2}$ (C) $10\sqrt{3}$ (D) 20

(E) It cannot be determined from the information given.

DIF: -1.05



Note: Figure not drawn to scale.

18. An X is drawn on a billboard with the dimensions shown in the figure above. If the 2 segments forming the X have equal lengths and bisect each other, what is the value of q?

(A) 45
(B) 60
(C) 75
(D) 90
(E) It cannot be determined from the information given.

DIF: -1.01

14. What is the greatest possible value of $\frac{x+y+z}{xyz}$ if x, y, and z represent different positive one-digit integers?

(A) $\frac{1}{21}$

(B) $\frac{1}{6}$

(C) 1

(D) $\frac{329}{168}$

(E) 6

DIF: -1.10

5. Car X averages 18 miles per gallon in city driving and 27 miles per gallon in highway driving. If the fuel tank capacity is 25 gallons, how many more miles per tankful of fuel can be expected in highway driving than in city driving?

(A) 562.5 (B) 225 (C) 34 (D) 22.5 (E) 9

DIF: -1.85
DIF: -1.28 (Black/White)

Appendix C

Table C1

Means and Standard Deviations of Categories for Variables on GMAT Verbal Tests

Variable/Categories	Number of Items	Comparison			
		F/M		B/W	
		Mean DIF	SD	Mean DIF	SD
Item Type					
Reading comprehension	74	-.14	.41	-.02	.44
Sentence correction	74	.18	.42	.03	.35
TOTAL		.02	.44	.00	.40
SIGNIFICANCE		.00		.42	

Table C2

Means and Standard Deviations of Categories for Variables on GMAT Verbal Tests: Reading Comprehension

Variable/Categories*	Number of Items	Comparison			
		F/M		B/W	
		Mean DIF	SD	Mean DIF	SD
Minority Reference - Stimulus					
Black	8	-.14	.39	.43	.50
Hispanic	9	-.32	.34	.52	.28
Other	0	-	-	-	-
General	57	-.11	.41	-.17	.33
Nothing	0	-	-	-	-
TOTAL	74	-.14	.41	-.02	.44
SIGNIFICANCE		.34		.00	
Gender Reference - Stimulus					
Female	0	-	-	-	-
Male	8	-.19	.22	-.24	.21
Mixed	66	-.13	.42	.00	.46
Neutral	0	-	-	-	-
TOTAL	74	-.14	.41	-.02	.44
SIGNIFICANCE		.72		.16	

* Categories that contain no items were dropped before the F test was performed; thus, the F test analyzes differences in the means of the categories that contain items.

(continued)

Table C2 (continued)

Means and Standard Deviations of Categories for Variables on GMAT Verbal Tests:

Reading Comprehension

Variable/Categories	Number of Items	Comparison			
		F/M		B/W	
		Mean DIF	SD	Mean DIF	SD
Negative Item					
Negative Stem	7	-.12	.55	.19	.49
Positive Stem	67	-.14	.39	-.05	.43
TOTAL	74	-.14	.41	-.02	.44
SIGNIFICANCE		.91		.18	
Roman Numeral Format					
Roman	4	-.17	.43	-.22	.44
Non-Roman	70	-.14	.41	-.01	.44
TOTAL	74	-.14	.41	-.02	.44
SIGNIFICANCE		.88		.38	
Number of Words in Stimulus (Passage)					
Less than 450 Words	17	.04	.39	-.13	.28
450 - 470 Words	43	-.22	.42	-.05	.45
More than 470 Words	14	-.13	.33	.18	.52
TOTAL	74	-.14	.41	-.02	.44
SIGNIFICANCE		.11		.13	

(continued)

Table C2 (continued)

Means and Standard Deviations of Categories for Variables on GMAT Verbal Tests:

Reading Comprehension

Variable/Categories	Number of Items	Comparison			
		F/M		B/W	
		Mean DIF	SD	Mean DIF	SD
Subject Content I					
Humanities	26	-.21	.39	-.19	.33
Social Studies	26	-.07	.44	.30	.44
Science	22	-.14	.38	-.21	.33
Business Related	0	-	-	-	-
TOTAL	74	-.14	.41	-.02	.44
SIGNIFICANCE		.52		.00	
Question Format					
Open Stem	47	-.14	.43	.01	.41
Closed Stem	27	-.13	.37	-.08	.49
TOTAL	74	-.14	.41	-.02	.44
SIGNIFICANCE		.90		.41	
Length of Stem					
Less than 15 Words	38	-.25	.39	-.10	.37
15 - 24 Words	27	.02	.40	.02	.45
25 - 34 Words	6	-.22	.28	.34	.59
More than 34 Words	3	-.03	.42	-.20	.50
TOTAL	74	-.14	.41	-.02	.44
SIGNIFICANCE		.06		.13	

(continued)

Table C2 (continued)

Means and Standard Deviations of Categories for Variables on GMAT Verbal Tests:

Reading Comprehension

Variable/Categories	Number of Items	Comparison			
		F/M		B/W	
		Mean DIF	SD	Mean DIF	SD
Line References					
Stem Directs Candidate to a Specific Line	12	-.15	.36	.06	.38
Stem Directs Candidate to a Specific Paragraph	1	-.06	-	.08	-
Stem Does Not Direct Candidate to Specific Line or Paragraph	61	-.14	.42	-.04	.46
TOTAL	74	-.14	.41	-.02	.44
SIGNIFICANCE		.98		.75	
Length of Options					
Less than 20 Words	20	-.23	.46	-.12	.43
20 - 39 Words	13	-.27	.31	-.03	.42
40 - 59 Words	22	-.12	.41	-.05	.39
60 - 79 Words	9	.04	.31	.18	.54
80 - 99 Words	8	.04	.39	.11	.45
100 Words or Over	2	-.10	.24	-.23	.22
TOTAL	74	-.14	.41	-.02	.44
SIGNIFICANCE		.37		.56	

(continued)

Table C2 (continued)

Means and Standard Deviations of Categories for Variables on GMAT Verbal Tests:

Reading Comprehension

Variable/Categories	Number of Items	Comparison			
		F/M		B/W	
		Mean DIF	SD	Mean	SD DIF
Item Specifications					
Main Idea	13	-.09	.41	-.04	.33
Supporting Detail	15	-.08	.43	.10	.42
Inference	22	-.13	.45	-.17	.46
Application	7	-.24	.37	-.14	.33
Logical Structure	11	-.22	.33	.15	.50
Style	6	-.14	.29	.04	.43
TOTAL	74	-.14	.41	-.02	.44
SIGNIFICANCE		.94		.33	
Named II - Gender *					
Female	0	-	-	-	-
Male	40	-.18	.37	-.06	.44
Mixed or Unidentifiable	9	.24	.39	-.04	.29
TOTAL	49	-.10	.41	-.06	.42
SIGNIFICANCE		.01		.88	

* In 25 items, no specifically named person was included in the passage.

Table C3

Means and Standard Deviations of Categories for Variables on GMAT Verbal Tests:

Sentence Correction

Variable/Categories*	Number of Items	Comparison			
		F/M		B/W	
		Mean DIF	SD	Mean DIF	SD
Minority Reference – Stimulus					
Black	4	.35	.29	.07	.17
Hispanic	0	–	–	–	–
Other	4	.34	.59	.03	.17
General	41	.15	.46	.06	.37
Nothing	25	.17	.30	–.02	.36
TOTAL	74	.18	.42	.03	.35
SIGNIFICANCE		.70		.83	
Gender Reference – Stimulus					
Female	7	.29	.37	–.02	.39
Male	7	.57	.35	.15	.34
Mixed	35	.08	.47	.05	.33
Neutral	25	.17	.30	–.02	.36
TOTAL	74	.18	.42	.03	.35
SIGNIFICANCE		.04		.65	

* Categories that contain no items were dropped before the F test was performed; thus, the F test analyzes differences in the means of the categories that contain items.

(continued)

Table C3 (continued)

Means and Standard Deviations of Categories for Variables on GMAT Verbal Tests:

Sentence Correction

Variable/Categories	Number of Items	Comparison			
		F/M		B/W	
		Mean DIF	SD	Mean DIF	SD
Number of Words in Stimulus					
Less than 20 Words	11	-.06	.40	-.15	.35
20 - 29 Words	31	.27	.35	.08	.34
30 - 39 Words	27	.17	.48	.06	.35
40 Words and Above	5	.15	.33	-.06	.25
TOTAL	74	.18	.42	.03	.35
SIGNIFICANCE		.17		.27	
Subject Content I					
Business Related	14	.23	.38	-.11	.28
Domestic & Foreign Affairs	11	.05	.36	.11	.28
Humanities	7	.53	.36	.28	.29
Science	22	.11	.45	-.05	.38
Social Sciences	18	.18	.42	.06	.33
Other	2	.24	.04	-.05	.31
TOTAL	73	.18	.42	.02	.35
SIGNIFICANCE		.24		.17	

(continued)

Table C3 (continued)

Means and Standard Deviations of Categories for Variables on GMAT Verbal Tests:

Sentence Correction

Variable/Categories	Number of Items	Comparison			
		F/M		B/W	
		Mean DIF	SD	Mean DIF	SD
Subject Content II					
Literature Only	2	.84	.27	.13	.10
All Other Humanities	5	.41	.33	.34	.32
TOTAL	7	.53	.36	.28	.29
SIGNIFICANCE		.23		.47	
Underline Position					
Beginning of Sentence	9	.27	.41	.23	.39
Middle of Sentence	40	.17	.43	-.04	.34
End of Sentence	22	.14	.43	.06	.34
Whole Sentence	3	.32	.27	.09	.26
TOTAL	74	.18	.42	.03	.35
SIGNIFICANCE		.82		.21	
Underline Element					
Main Clause Only	38	.17	.39	-.02	.37
Subordinate Only	18	.19	.42	.05	.30
Combination - Main Clause & Subordinate Elements	18	.19	.47	.10	.35
TOTAL	74	.18	.42	.03	.35
SIGNIFICANCE		.98		.47	

(continued)

Table C3 (continued)

Means and Standard Deviations of Categories for Variables on GMAT Verbal Tests:

Sentence Correction

Variable/Categories	Number of Items	Comparison			
		F/M		B/W	
		Mean DIF	SD	Mean DIF	SD
Underline Length					
1 - 7 Words	24	.00	.48	-.04	.37
8 - 14 Words	30	.24	.34	.05	.35
15 & Over	20	.30	.39	.08	.32
TOTAL	74	.18	.42	.03	.35
SIGNIFICANCE		.04		.48	
Key					
Answer Choice A (Identical to Stem)	13	.22	.44	.14	.40
Answer Choices B - E	61	.17	.42	.00	.33
TOTAL	74	.18	.42	.03	.35
SIGNIFICANCE		.70		.20	

(continued)

Table C3 (continued)

Means and Standard Deviations of Categories for Variables on GMAT Verbal Tests:

Sentence Correction

Variable/Categories	Number of Items	Comparison			
		F/M		B/W	
		Mean DIF	SD	Mean DIF	SD
Major Testing Points					
Agreement	5	.55	.42	.04	.30
Construction - Grammatical	10	.22	.45	-.04	.43
Construction - Rhetorical	7	.04	.17	-.17	.36
Diction	1	.20	-	.65	-
Idiom	11	-.06	.61	.00	.33
Logical Predication	22	.15	.42	.10	.29
Negation	3	.15	.08	-.15	.56
Parallelism	4	.33	.13	-.19	.19
Verb Form	11	.30	.17	.17	.20
TOTAL	74	.18	.42	.03	.35
SIGNIFICANCE		.28		.22	

Table C4

Means and Standard Deviations of Categories for Variables on GMAT Quantitative Tests

Variable/Categories*	Number of Items	Comparison			
		F/M		B/W	
		Mean DIF	SD	Mean DIF	SD
Minority Reference - Stimulus					
Black	0	-	-	-	-
Hispanic	2	-.08	.35	-.73	.05
Other	0	-	-	-	-
General	47	-.10	.51	-.10	.35
Nothing	146	.05	.55	.10	.53
TOTAL	195	.01	.55	.05	.50
SIGNIFICANCE		.24		.01	
Gender Reference - Stimulus					
Female	6	.17	.40	-.47	.16
Male	6	-.23	.37	-.01	.35
Mixed	37	-.13	.52	-.08	.36
Neutral	146	.05	.55	.10	.53
TOTAL	195	.01	.55	.05	.50
SIGNIFICANCE		.19		.01	

* Categories that contain no items were dropped before the F test was performed; thus, the F test analyzes differences in the means of the categories that contain items.

(continued)

Table C4 (continued)

Means and Standard Deviations of Categories for Variables on GMAT Quantitative Tests

Variable/Categories	Number of Items	Comparison			
		F/M		B/W	
		Mean DIF	SD	Mean DIF	SD
Item Type					
Data Sufficiency	75	.16	.50	.02	.45
Problem Solving	120	-.08	.55	.06	.53
TOTAL	195	.01	.55	.05	.50
SIGNIFICANCE		.00		.51	
Real/Pure					
Real - Includes Story or Word Problems	105	-.15	.58	-.10	.44
Pure - Theory Includes Formulae & Calculations	90	.19	.44	.21	.52
TOTAL	195	.01	.55	.05	.50
SIGNIFICANCE		.00		.00	
Primary Content Area					
Arithmetic	101	.01	.61	-.05	.44
Algebra	59	.10	.45	.27	.54
Geometry	35	-.13	.48	-.06	.46
TOTAL	195	.01	.55	.05	.50
SIGNIFICANCE		.13		.00	

(continued)

Table C4 (continued)

Means and Standard Deviations of Categories for Variables on GMAT Quantitative Tests

Variable/Categories	Number of Items	Comparison			
		F/M		B/W	
		Mean DIF	SD	Mean DIF	SD
Multiple Categories					
Single Only	167	.01	.54	.05	.48
Arithmetic/ algebra	19	.01	.65	.12	.56
Arithmetic/ geometry	4	-.18	.34	-.50	.41
Algebra/ geometry	5	.09	.27	-.06	.65
TOTAL	195	.01	.55	.05	.50
SIGNIFICANCE		.90		.14	
Stimulus Format I: Pictures					
Figure	24	-.10	.55	-.08	.49
Graph	3	-.12	.15	-.28	.13
Table	5	.06	.53	-.10	.22
None	163	.03	.55	.07	.51
TOTAL	195	.01	.55	.05	.50
SIGNIFICANCE		.74		.31	

(continued)

Table C4 (continued)

Means and Standard Deviations of Categories for Variables on GMAT Quantitative Tests

Variable/Categories	Number of Items	Comparison			
		F/M		B/W	
		Mean DIF	SD	Mean DIF	SD
Stimulus Format II: Letters for Mathematical Quantities					
In option	0	-	-	-	-
In stem	31	.12	.49	.36	.53
In both	28	.18	.34	.36	.42
No letters	136	-.05	.58	-.09	.44
TOTAL	195	.01	.55	.05	.50
SIGNIFICANCE		.06		.00	
Subject Content I					
Business related	50	-.05	.42	-.07	.35
Other	145	.03	.58	.09	.54
TOTAL	195	.01	.55	.05	.50
SIGNIFICANCE		.37		.05	

Table C4 (continued)

Means and Standard Deviations of Categories for Variables on GMAT Quantitative Tests

Variable/Categories	Number of Items	Comparison			
		F/M		B/W	
		Mean DIF	SD	Mean DIF	SD
Quantitative Content Area					
Algebraic manipulations	42	.14	.37	.25	.56
Ratio/proportion	21	-.21	.50	.08	.54
Factor/multiple	18	.29	.37	.03	.43
None of the above	114	-.04	.60	-.03	.46
TOTAL	195	.01	.55	.05	.50
SIGNIFICANCE		.01		.02	

Table C5

Means and Standard Deviations of Categories for Variables on GMAT Quantitative Test:

Problem Solving

Variable/Categories*	Number of Items	Comparison			
		F/M		B/W	
		Mean DIF	SD	Mean DIF	SD
Negative Item					
Negative Stem	4	.61	.33	.08	.52
Positive Stem	116	-.10	.54	.06	.53
TOTAL	120	-.08	.55	.06	.53
SIGNIFICANCE		.01		.94	
Roman Numeral Format					
Roman	3	-.03	.18	.34	.03
Non-Roman	117	-.08	.56	.06	.53
TOTAL	120	-.08	.55	.06	.53
SIGNIFICANCE		.89		.36	
Number of Words in Stem					
Below 50 Words	90	-.07	.51	.14	.54
50 - 99 Words	28	-.18	.65	-.17	.43
100 - 149 Words	0	—	—	—	—
150 - 199 Words	2	.67	.15	.02	.02
200 and over	0	—	—	—	—
TOTAL	120	-.08	.55	.06	.53
SIGNIFICANCE		.10		.02	

*Categories that contain no items were dropped before the F test was performed; thus, the F test analyzes differences in the means of the categories that contain items.

Table C6

Means and Standard Deviations of Categories for Variables on GMAT Quantitative Test:

Data Sufficiency

Variable/Categories*	Number of Items	Comparison			
		F/M		B/W	
		Mean DIF	SD	Mean	SD DIF
Data Sufficiency Key					
Option A	16	.20	.51	-.07	.46
Option B	15	.15	.47	-.12	.41
Option C	15	.01	.38	.08	.38
Option D	15	.15	.57	.06	.48
Option E	14	.28	.52	.14	.45
TOTAL	75	.16	.50	.02	.45
SIGNIFICANCE		.70		.49	
Number of words in stimulus					
Below 50	53	.18	.52	.09	.43
50 - 99	22	.11	.49	-.15	.45
100 - 149	0	—	—	—	—
150 - 199	0	—	—	—	—
200 and over	0	—	—	—	—
TOTAL	75	.16	.50	.02	.45
SIGNIFICANCE		.63		.04	

* Categories that contain no items were dropped before the F test was thus, the F test analyzes differences in the means of the categories that contain items.

SAMPLE SENTENCE CORRECTION ITEMS

Shorter Underline

16. The insulin pump maintains the blood sugar level of diabetics as well if not better than periodic injections.

(A) as well if not better than
(B) at least as well as
(C) as well or better than
(D) equally well if not better than
(E) perhaps better but at least as well as

DIF: -.98 (Female/Male Analysis)

Larger Underline

12. Elephant seals, once near extinction after hunters slaughtered them indiscriminately in the last century, have now become such prolific breeders that they are colonizing the beaches of California.

(A) Elephant seals, once near extinction after hunters slaughtered them indiscriminately in the last century, have now become such prolific breeders that they are
(B) Elephant seals, once near extinction after hunters slaughtered them indiscriminately in the last century, have now become such a prolific breeder that it is
(C) The elephant seal, once near extinction after hunters slaughtered them indiscriminately in the last century, have now become such prolific breeders that they are
(D) The elephant seal, once near extinction after hunters slaughtered it indiscriminately in the last century, have now become such prolific breeders that it is
(E) Once near extinction after hunters slaughtered them indiscriminately, the elephant seal has now become such prolific breeders that they are

DIF: +1.13 (Female/Male Analysis)

BEST COPY AVAILABLE